Focal and Global Knowledge Distillation for Detectors

Zhendong Yang[*1,2]   Zhe Li[2]   Xiaohu Jiang[1]   Yuan Gong[1]
Zehuan Yuan[2]   Danpei Zhao[3]   Chun Yuan[†1]

[1]Tsinghua Shenzhen International Graduate School   [2]ByteDance Inc
[3]BeiHang University

{yangzd21,jiangxh21,gong-y21}@mails.tsinghua.edu.cn   {lizhe.axel,yuanzehuan}@bytedance.com
zhaodanpei@buaa.edu.cn   yuanc@sz.tsinghua.edu.cn

Abstract

Knowledge distillation has been applied to image classification successfully. However, object detection is much more sophisticated and most knowledge distillation methods have failed on it. In this paper, we point out that in object detection, the features of the teacher and student vary greatly in different areas, especially in the foreground and background. If we distill them equally, the uneven differences between feature maps will negatively affect the distillation. Thus, we propose Focal and Global Distillation (FGD). Focal distillation separates the foreground and background, forcing the student to focus on the teacher's critical pixels and channels. Global distillation rebuilds the relation between different pixels and transfers it from teachers to students, compensating for missing global information in focal distillation. As our method only needs to calculate the loss on the feature map, FGD can be applied to various detectors. We experiment on various detectors with different backbones and the results show that the student detector achieves excellent mAP improvement. For example, ResNet-50 based RetinaNet, Faster RCNN, RepPoints and Mask RCNN with our distillation method achieve 40.7%, 42.0%, 42.0% and 42.1% mAP on COCO2017, which are 3.3, 3.6, 3.4 and 2.9 higher than the baseline, respectively. Our codes are available at https://github.com/yzd-v/FGD.

1. Introduction

Recently, deep learning has achieved great success in various domains [8, 9, 22, 24]. To get better performance, we usually use a larger backbone, which needs more compute resources and inferences more slowly. To get over this, knowledge distillation has been proposed [11]. Knowl-

*This work was performed while Zhendong worked as an intern at ByteDance.
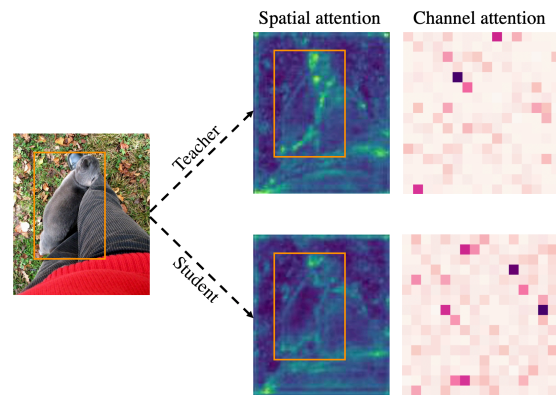
†Corresponding author

Figure 1. Visualization of the spatial and channel attention map from the teacher detector (RetinaNet-ResNeXt101) and the student detector (RetinaNet-ResNet50).

| | distillation area | | | mAP | mAR |
	fg	bg	split		
RetinaNet Res101-Res50	×	×		37.4	53.9
	✓	×		39.3	55.6
	×	✓		39.2	55.8
	✓	✓	×	38.9	55.1
	✓	✓	✓	**39.4**	**56.1**

Table 1. Comparisons of different distillation areas. **fg**: foreground. **bg**: background. **split**: split the foreground and background and distill them with different weights.

edge distillation is a method to inherit the information from a large teacher network to a compact student network and achieve strong performance without extra cost during inference time. However, most distillation methods [10, 27, 33, 34] are designed for image classification, which lead to trivial improvements for object detection.

It is well acknowledged that the extreme foreground-

background class imbalance is a key point in object detection [17]. The imbalanced ratio also harms the distillation for object detection. There are some efforts for this problem. Chen *et al.* [3] distributes a weight to suppress the background. Mimick [15] distills the positive area proposed by region proposal network of the student. FGFI [28] and TADF [25] use the fine-grained and Gaussian Mask to select the distillation area, respectively. Defeat [7] distills the foreground and background separately. However, where is the key area for distillation is still not clear.

In order to explore the difference between the features of students and teachers, we do the visualization of the spatial and channel attention. As the Fig. 1 shows, the difference between student's attention and teacher's attention in the foreground is quite significant, while that in the background is relatively small. This may lead to different difficulties in learning the foreground and background. In this paper, we further explore the influence of the foreground and background in knowledge distillation on object detection. We design experiments by decoupling the foreground and background in the distillation. Surprisingly, as shown in Tab. 1, the performance of distillation on the foreground and background together is the worst, even worse than only using foreground or background. This phenomenon suggests that the uneven differences in the feature map can negatively affect distillation. Besides, as shown in Fig. 1, the attention between each channel is also very different. Thinking one step deeper, not only are there negative influences between the foreground and the background, but also between the pixels and the channels. Therefore, we propose focal distillation. While separating the foreground and background, focal distillation also calculates the attention of different pixels and channels in teacher's feature, allowing the student to focus on teacher's crucial pixels and channels.

However, just focusing on key information is not enough. It is well known that global context also plays an important role in detection. A lot of relation modules have been successfully applied into detection, such as non-local [29], GcBlock [2], relation network [12], which have greatly improved the performance of detectors. In order to compensate for the missing global information in focal distillation, we further propose global distillation. In global distillation, we utilize GcBlock to extract the relation between different pixels and then distill them from teachers to students.

As we analyzed above, we propose **F**ocal and **G**lobal **D**istillation (FGD), combining focal distillation and global distillation, as shown in Fig. 2. All loss functions are only calculated on features, so that FGD can be used directly on various detectors, including two-stage models, anchor-based one-stage models and anchor-free one-stage models. Without bells and whistles, we achieve state-of-the-art performances in object detection with FGD. In a nutshell, the contributions of this paper are:

- We present that the pixels and channels that teacher and student pay attention to are quite different. If we distill the pixels and channels without distinguishing them, it will result in a trivial improvement.

- We propose focal and global distillation, which enables the student not only to focus on the teacher's critical pixels and channels, but also to learn the relation between pixels.

- We verify the effectiveness of our method on various detectors via extensive experiments on the COCO [18], including one-stage, two-stage, anchor-free methods, achieving state-of-the-art performance.

2. Related Work

2.1. Object Detection

Object detection is a fundamental and challenging task in computer vision. The CNN-based detection networks with high performance are divided into two-stage [1, 8, 22], anchor-based one-stage [17, 19, 21] and anchor-free one-stage detectors [6, 26, 32]. One-stage detectors get the classification and bounding box of targets on feature maps directly. In contrast, two-stage detectors utilize RPN and RCNN head to achieve better results but cost more time. Prior anchor boxes provide one-stage models with proposals to detect targets. However, the number of anchor boxes is far more than targets, which brings extra computation. While anchor-free detectors show a way to predict the key point and location of targets directly. Although there are different detection heads, their inputs are all features. Therefore, our feature-based knowledge distillation method can be applied in almost all detectors.

2.2. Knowledge Distillation

Knowledge distillation is a method of model compression without changing the network structure. It is first proposed by Hinton *et al.* [11], which uses the output as soft labels to transfer the dark knowledge from a large teacher network to a small student network for the classification task. Moreover, FitNet [23] proves that the semantic information from intermediate is also helpful to guide the student model. There have been many works [10, 27, 33, 34] that improve the student classifiers significantly.

Recently, some works have successfully applied knowledge distillation to detectors. Chen *et al.* [3] first apply knowledge distillation to detection by distilling knowledge on the neck feature, the classification head, and the regression head. Nevertheless, distilling the whole feature may introduce much noise because of the imbalance between the foreground and background. Li *et al.* [15] choose the features sampled from RPN to calculate distillation loss.
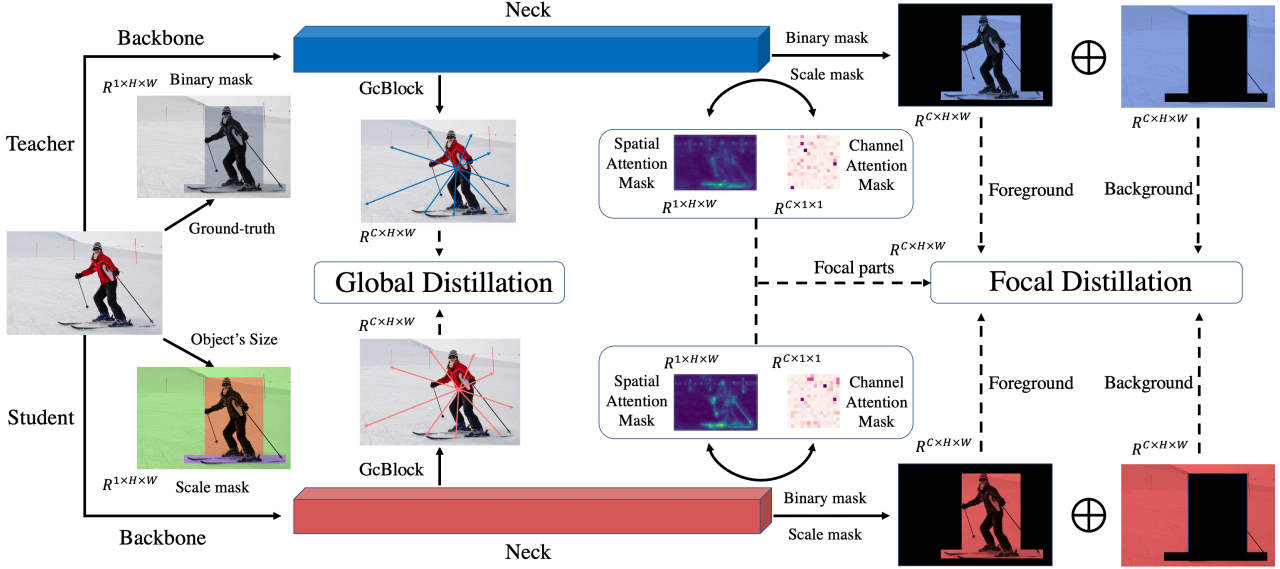
Figure 2. An illustration of FGD, including focal distillation and global distillation. Focal distillation not only separates the foreground and the background, but also enables the student network to better pay attention to the important information in the teacher network's feature map. Global distillation bridges the gap between the global context of the student and the teacher.

Wang *et al.* [28] propose the fine-grained mask to distill the regions calculated by ground-truth bounding boxes. Sun *et al.* [25] utilize the Gaussian Mask to cover the ground-truth for distillation. Such methods lack the distillation for the background. Without distinguishing the foreground and background, GID [5] distills the areas where the performance of the student and teacher is different. Guo *et al.* [7] shows that both the foreground and background play important roles for distillation, and distilling them separately benefits the student more. Both of their methods distill the knowledge from the background and get significant results. However, they treat all the pixels and channels equally. FKD [35] uses attention masks and Non-local module [29] to guide the student and distills the relation, respectively. However, it distills the foreground and background together.

The critical problem of distillation for detection is to select the valuable area for distillation. The previous distillation methods treat all the pixels and channels equally [5, 7, 25, 28] or distill all the areas [35] together. Most methods lack the distillation of the global context information. In this paper, we use ground-truth boxes to separate the images, and then use the attention masks from the teacher to select crucial parts for distillation. In addition, we capture the global relations between different pixels and distill them to the student, which brings another improvement.

3. Method

Most detectors have used FPN [16] to utilize the multi-scale semantic information. The features from FPN fuse different levels of semantic information from the backbone and are used to predict directly. Transferring the knowledge of these features from the teacher has significantly improved the performance of the student. Generally, the distillation of the features can be formulated as:

$$L_{fea} = \frac{1}{CHW} \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(F_{k,i,j}^{T} - f(F_{k,i,j}^{S}) \right)^2 \quad (1)$$

where F^T and F^S denote the feature from the teacher and student, respectively, and f is the adaptation layer to reshape the F^S to the same dimension as F^T. H, W denote the height and width of the feature and C is the channel.

However, such methods treat all the parts equally and lack the distillation of the global relations between different pixels. To get over the above problems, we propose FGD, which includes focal and global distillation, as shown in Fig. 2. Here we will introduce our method in detail.

3.1. Focal Distillation

For the foreground and background imbalance, we propose focal distillation to separate the images and guide the student to focus on crucial pixels and channels. The comparison of the distillation areas can be seen in Fig. 3.

Firstly we set a binary mask M to separate the background and foreground:

$$M_{i,j} = \begin{cases} 1, & \text{if } (i,j) \in r \\ 0, & \text{Otherwise} \end{cases} \quad (2)$$

where r denotes the ground-truth boxes and i, j are the the horizontal and vertical coordinates of the feature map, respectively. If (i, j) falls in the ground truth, then $M_{i,j} = 1$, otherwise it is 0.

The targets with larger-scale will occupy more loss because they own more pixels, which will influence the distillation of the small targets. And the ratios of foreground to background vary greatly in different images. Therefore, in order to treat different targets equally and balance the loss of foreground and background, we set a scale mask S as:

$$S_{i,j} = \begin{cases} \frac{1}{H_r W_r}, & \text{if } (i,j) \in r \\ \frac{1}{N_{bg}}, & \text{Otherwise} \end{cases} \quad (3)$$

$$N_{bg} = \sum_{i=1}^{H} \sum_{j=1}^{W} (1 - M_{i,j}) \quad (4)$$

where H_r and W_r denote the height and width of the ground-truth box r. If a pixel belongs to different targets, we choose the smallest box to calculate the S.

SENet [13] and CBAM [30] show that focusing on crucial pixels and channels helps CNN-based models get better results. Zagoruyko *et al.* [34] use a simple way to get the spatial attention mask and improve the performance of distillation. In this paper, we apply a similar method to select focal pixels and channels, and then get corresponding attention masks. We calculate the absolute mean values on different pixels and different channels, respectively:

$$G^S(F) = \frac{1}{C} \cdot \sum_{c=1}^{C} |F_c| \quad (5)$$

$$G^C(F) = \frac{1}{HW} \cdot \sum_{i=1}^{H} \sum_{j=1}^{W} |F_{i,j}| \quad (6)$$

where H, W, C denote the feature's height, width, and channel. G^S and G^C are the spatial and channel attention map. Then the attention mask can be formulated as:

$$A^S(F) = H \cdot W \cdot softmax(G^S(F)/T) \quad (7)$$

$$A^C(F) = C \cdot softmax(G^C(F)/T) \quad (8)$$

where T is the temperature hyper-parameter proposed by Hinton *et al.* [11] to adjust the distribution.

There are significant differences between the masks of the student and teacher. In training process, we use the teacher's masks to guide the student. With the binary mask M, scale mask S, attention mask A^S and A^C, we propose

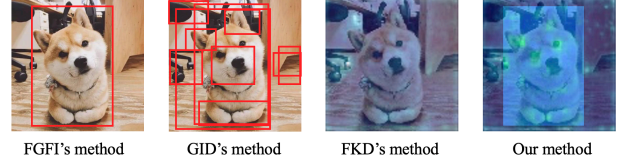| FGFI's method | GID's method | FKD's method | Our method |

Figure 3. Comparison of the distillation areas between our method (FGD) and other methods. FGFI and GID only distill the areas in the red bounding box. The areas where GID and FKD distill are changeable during training. Different colors mean different weights and the green parts mean the spatial attention pixels.

the feature loss L_{fea} as follows:

$$L_{fea} = \alpha \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} M_{i,j} S_{i,j} A_{i,j}^S A_k^C \left(F_{k,i,j}^T - f(F_{k,i,j}^S) \right)^2$$
$$+ \beta \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} (1 - M_{i,j}) S_{i,j} A_{i,j}^S A_k^C \left(F_{k,i,j}^T - f(F_{k,i,j}^S) \right)^2 \quad (9)$$

where A^S and A^C denote the spatial and channel attention mask of the teacher detector, respectively. F^T and F^S denote the feature maps of the teacher detector and student detector, respectively. α and β are the hyper-parameters to balance the loss between foreground and background.

Besides, we use attention loss L_{at} to force the student detector to mimic the spatial and channel attention mask of the teacher detector, which is formulated as:

$$L_{at} = \gamma \cdot \left(l(A_t^S, A_S^S) + l(A_t^C, A_S^C) \right) \quad (10)$$

where t and s denote the teacher and student. l denotes L1 loss and γ is a hyper-parameter to balance the loss.

The focal loss L_{focal} is the sum of feature loss L_{at} and attention loss L_{at}:

$$L_{focal} = L_{fea} + L_{at} \quad (11)$$

3.2. Global Distillation

The relation [2,12,29] between different pixels has valuable knowledge and is utilized to improve the performance for detection tasks. And in Sec. 3.1, we utilize Focal Distillation to separate the images and force the student focus on crucial parts. However, such distillation cuts off the relation between foreground and background. So here we propose Global Distillation, which aims to extract the global relation between different pixels from the feature maps and distill it from the teacher to the student.

As shown in Fig. 4, we utilize GcBlock [2] to capture the global relation information in a single image and force
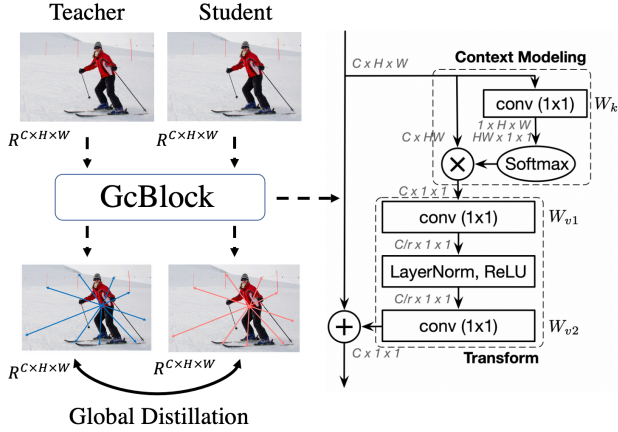
Figure 4. The Global Distillation with GcBlock. The inputs are the feature maps from the teacher's neck and student's neck, respectively.

the student detector to learn the relation from the teacher detector. The global loss L_{global} is as follows:

$$
\begin{aligned}
L_{global} =& \lambda \cdot \sum \left(\mathcal{R}(F^T) - \mathcal{R}(F^S) \right)^2 \\
\mathcal{R}(F) =& F + W_{v2}(ReLU(LN(W_{v1} \\
& (\sum_{j=1}^{N_p} \frac{e^{W_k F_j}}{\sum_{m=1}^{N_p} e^{W_k F_M}} F_j))))
\end{aligned}
\tag{12}
$$

where W_k, W_{v1} and W_{v2} denote convolutional layers, LN denotes the layer normalization, N_p is the number of pixels in the feature and λ is a hyper-parameter to balance the loss.

3.3. Overall loss

To sum up, we train the student detector with the total loss as follows:

$$
L = L_{original} + L_{focal} + L_{global}
\tag{13}
$$

where $L_{original}$ is the original loss for detectors.

The distillation loss is calculated just on feature maps, which can be obtained from the neck of the detectors. So it can be easily applied to different detectors.

4. Experiments

4.1. Dataset

We evaluate our knowledge distillation method on COCO dataset [18], which contains 80 object classes. We use the 120k train images for training and 5k val images for testing for all the experiments. The performances of different detectors are evaluated in Average Precision and Average Recall.

Method	mAP	AP_S	AP_M	AP_L
RetinaNet-Res101(T)	38.9	21.0	42.8	52.4
RetinaNet-Res50(S)	37.4	20.6	40.7	49.7
FGFI [28]	38.6	21.4	42.5	51.5
GID [5]	39.1	22.8	43.1	52.3
Ours	39.6	22.9	43.7	53.6
Ours †	39.7	22.0	43.7	53.6
RCNN-Res101(T)	39.8	22.5	43.6	52.8
RCNN-Res50(S)	38.4	21.5	42.1	50.3
FGFI [28]	39.3	22.5	42.3	52.2
GID [5]	40.2	22.7	44.0	53.2
Ours	40.4	22.8	44.5	53.5
Ours †	40.5	22.6	44.7	53.2
FCOS-Res101(T)	40.8	24.2	44.3	52.4
FCOS-Res50(S)	38.5	21.9	42.8	48.6
GID [5]	42.0	25.6	45.8	54.2
Ours	42.1	27.0	46.0	54.6
Ours †	42.7	27.2	46.5	55.5

Table 2. Results of different distillation methods with different detection frameworks on COCO dataset. **T** and **S** mean the teacher and student detector, respectively. FGFI can only be applied to an anchor-based detector. † means using inheriting strategy. We train the FCOS with tricks including GIoULoss, norm-on-bbox and center-sampling which is the same as GID.

4.2. Details

We conduct experiments on different detection frameworks, including two-stage models [22], anchor-based one-stage models [17], and anchor-free one-stage models [26, 32]. Besides, we verify our method on the Mask RCNN [8] and get significant improvement for instance segmentation. Kang *et al.* [14] propose inheriting strategy which initializes the student with the teacher's neck and head parameters and gets better results. Here we use this strategy to initialize the student which has the same head structure as the teacher. All the experiments are conducted with mmdetection [4] with Pytorch [20].

FGD uses $\alpha, \beta, \gamma, \lambda$ to balance the loss of foreground and background in Eq. (9), attention loss in Eq. (10) and global loss in Eq. (12), respectively. And $T = 0.5$ is used to adjust the attention distribution for all the experiments. We adopt the hyper-parameters $\{\alpha = 5 \times 10^{-5}, \beta = 2.5 \times 10^{-5}, \gamma = 5 \times 10^{-5}, \lambda = 5 \times 10^{-7}\}$ for all the two-stage models, $\{\alpha = 1 \times 10^{-3}, \beta = 5 \times 10^{-4}, \gamma = 1 \times 10^{-3}, \lambda = 5 \times 10^{-6}\}$ for all the anchor-based one-stage models, $\{\alpha = 1.6 \times 10^{-3}, \beta = 8 \times 10^{-4}, \gamma = 8 \times 10^{-3}, \lambda = 8 \times 10^{-6}\}$ for all the anchor-free one-stage models. We train all the detectors for 24 epochs with SGD optimizer, which the momentum is 0.9 and the weight decay is 0.0001.

Teacher	Student	mAP	AP_S	AP_M	AP_L	mAR	AR_S	AR_M	AR_L
RetinaNet ResNeXt101	RetinaNet-Res50	37.4	20.6	40.7	49.7	53.9	33.1	57.7	70.2
	FKD [35]	39.6(+2.2)	22.7	43.3	52.5	56.1(+2.2)	36.8	60.0	72.1
	Ours	40.4(+3.0)	23.4	44.7	54.1	56.7(+2.8)	37.6	61.5	72.4
	Ours†	40.7(+3.3)	22.9	45.0	54.7	56.8(+2.9)	36.5	61.4	72.8
Cascade Mask RCNN ResNeXt101	Faster RCNN-Res50	38.4	21.5	42.1	50.3	52.0	32.6	55.8	66.1
	FKD [35]	41.5(+3.1)	23.5	45.0	55.3	54.4(+2.4)	34.0	58.2	69.9
	Ours	42.0(+3.6)	23.8	46.4	55.5	55.4(+3.4)	35.5	60.0	70.0
RepPoints ResNeXt101	RepPoints-Res50	38.6	22.5	42.2	50.4	55.1	34.9	59.4	70.3
	FKD [35]	40.6(+2.0)	23.4	44.6	53.0	56.9(+1.8)	37.3	60.9	71.4
	Ours	41.3(+2.7)	24.5	45.2	54.0	58.4(+3.3)	39.1	62.9	74.2
	Ours†	42.0(+3.4)	24.0	45.7	55.6	58.2(+3.1)	37.8	62.2	73.3

Teacher	Student	Boundingbox AP				Mask AP			
		mAP	AP_S	AP_M	AP_L	mAP	AP_S	AP_M	AP_L
Cascade Mask RCNN ResNeXt101	Mask RCNN-Res50	39.2	22.9	42.6	51.2	35.4	19.1	38.6	48.4
	FKD [35]	41.7(+2.5)	23.4	45.3	55.8	37.4(+2.0)	19.7	40.5	52.1
	Ours	42.1(+2.9)	23.7	46.2	55.7	37.8(+2.4)	19.7	41.3	52.3

Table 3. Results of more detectors with stronger teacher detectors on COCO dataset. † means using inheriting strategy, which can only be applied when the student and teacher have the same head structure.

4.3. Main Results

Our method can be applied to different detection frameworks easily, so we first conduct experiments on three popular detectors, including a two-stage detector (Faster RCNN), an anchor-based one-stage detector (RetinaNet) and an anchor-free detector (FCOS). We compare with other two knowledge distillation methods [5, 28] for object detection. In the experiments, we choose the detectors with ResNet-50 [9] as the students and the identical detectors with ResNet-101 as the teachers. As shown in Tab. 2, our distillation method surpasses the other two state-of-the-art methods. All the student detectors gain significant AP improvements with the knowledge transferred from teacher detectors, *e.g.* the RetinaNet based ResNet-50 gets 2.3 mAP improvement on COCO dataset. Furthermore, in this Res101-Res50 setting, the student detectors even outperform the teacher detectors by training with FGD.

4.4. Distillation of more detectors with stronger students and teachers

Our method can also be applied between heterogeneous backbones, *e.g.* the ResNeXt [31] based teacher detector distill the ResNet based student detector. Here we conduct experiments on more detectors and use stronger backbone-based teacher detectors. And we compare the results with FKD [28], which is another effective and general distillation method. As shown in Tab. 3, all the student detectors achieve significant improvements on both AP and AR.

Besides, comparing the results with Tab. 2, we find that student detectors perform better with stronger teacher detectors, *e.g.* Retina-Res50 model achieves 40.7 and 39.7 mAP with ResNeXt101 and ResNet101 based teacher, respectively. The comparisons show that student detectors get the better feature by mimicking the feature maps of stronger backbones-based teacher detectors.

FGD only needs to calculate the distillation loss on the feature maps. So we also apply our method to Mask RCCN for object detection and instance segmentation. And in this experiment, we use the bounding box labels for the focal distillation. As shown in Tab. 3, our method brings 2.9 Boundingbox AP gains and 2.4 Mask AP gains, which proves our distillation method is also effective for instance segmentation.

4.5. Better feature with FGD

As shown in Tab. 2 and Tab. 3, initializing the student with the teacher's neck and head parameters brings another improvement, which indicates the student gets a similar feature with the teacher. So in this subsection, we visualize and compare the spatial attention mask and channel attention mask from teacher detector, student detector and student detector with FGD, which is shown in Fig. 5. Comparing the attention masks between the teacher and student, we can see they have a big difference in the distribution of pixels and channels before distillation, *e.g.* the teacher detector focuses more on the fingers and has a larger weight in channel 241. However, after training with FGD, the student de-
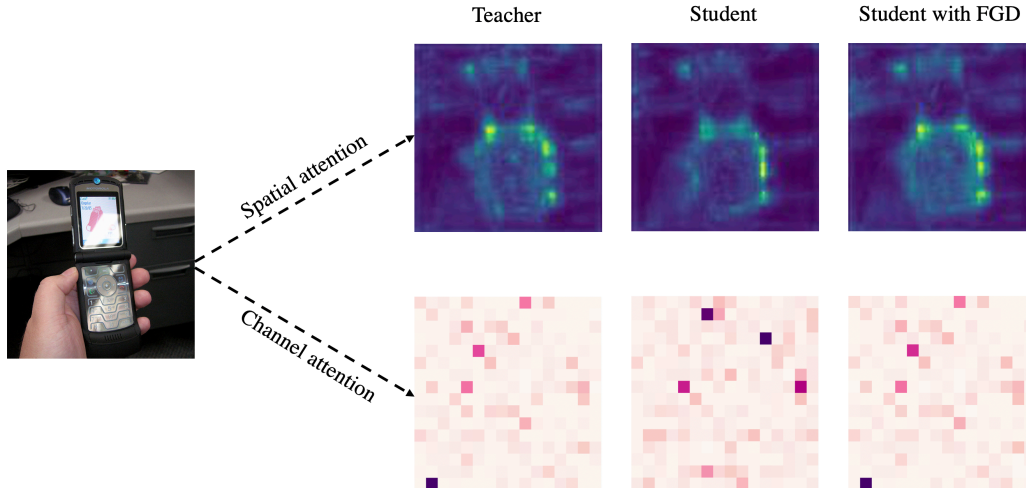
Figure 5. Visualization of the spatial and channel attention mask from different detectors. Each pixel in the channel attention mask means a channel. **Teacher detector**: RetinaNet-ResNeXt101. **Student detector**: RetinaNet-ResNet50

tector has a similar distribution of pixels and channels with the teacher detectors, which means the student focuses on the same parts as the teacher. This also explains how FGD helps the student detector perform better. Based on a similar feature, the student detector gets significant improvements and even outperforms the teacher detector.

4.6. Analysis

4.6.1 Sensitivity study of different losses

In this paper, we transfer the focal knowledge and global knowledge from the teacher to the student. In this subsection, we conduct experiments of focal loss (L_{focal}) and global loss (L_{global}) to investigate their influences on the student with RetinaNet. As shown in Tab. 4, both the focal loss and global loss lead to significant AP and AR improvements. Furthermore, considering targets with different sizes, we find L_{focal} benefits more to the large size targets and L_{global} benefits more to the small and medium targets. Besides, when combining L_{focal} and L_{global}, we achieve 40.4 mAP and 56.7 mAR, which indicates the focal loss and global loss are complementary to each other.

4.6.2 Sensitivity study of focal distillation

In focal distillation, we use the ground-truth boxes to separate the images and guide the student with the teacher's attention masks. In this subsection, we explore the effectiveness of focal distillation.

As shown in Tab. 1, we find distilling just on foreground or background both lead significant improvements. Here we analyze different error types to investigate their effectiveness, which is shown in Fig. 6. With the knowl-

Method	ReinaNet ResX101-Res50			
L_{focal}	-	✓	-	✓
L_{global}	-	-	✓	✓
mAP	37.4	40.2	40.2	**40.4**
AP_S	20.0	22.8	22.9	**23.4**
AP_M	40.7	44.0	44.3	**44.7**
AP_L	49.7	54.0	53.4	**54.1**
mAR	53.9	56.2	56.4	**56.7**
AR_S	33.1	36.8	37.3	**37.6**
AR_M	57.7	60.3	60.5	**61.5**
AR_L	70.2	72.3	72.2	**72.4**

Table 4. Ablation study of focal and global distillation.

edge from background, student detectors reduce the false-positive predictions and get higher mAP. In comparison, the foreground's distillation helps students detect more targets and reduce the false-negative predictions. In conclusion, the results show that both foreground and background are crucial and have different functions for the student detectors.

In this paper, we utilize the spatial and channel attention mask of the teacher to guide the student to focus on crucial parts. Here we conduct experiments with RetinaNet to explore the effects of each mask, which is shown in Tab. 5. Each attention mask improves the performance, especially the spatial attention mask which brings 2.6 mAP gains and 2.2 mAR gains. And the combination of two masks gets the best result. The experiments show both the attention masks help the student perform better.

(a) KD: foreground

(b) KD: background
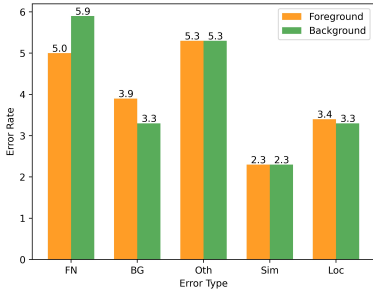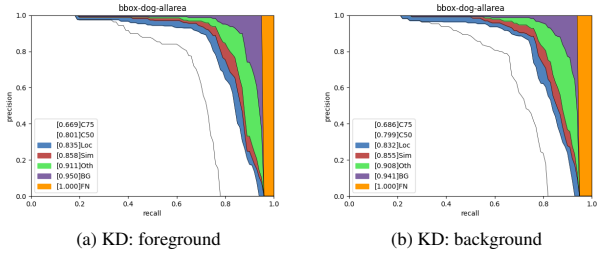

(c) Error of different types analysis

Figure 6. Different error types analyses of foreground and background distillation. **FN**: false negative prediction; **BG**: Background false positive prediction; **Oth**: classification errors; **Sim**: wrong class but correct supercategory; **Loc**: localization errors

Method	ReinaNet ResNeXt101-Res50			
Spatial attention	-	✓	-	✓
Channel attention	-	-	✓	✓
mAP	37.4	40.0	39.7	**40.2**
AP_S	20.0	22.3	22.0	**22.8**
AP_M	40.7	44.0	43.5	**44.0**
AP_L	49.7	53.6	53.4	**54.0**
mAR	53.9	56.1	55.8	**56.2**
AR_S	33.1	36.5	35.7	**36.8**
AR_M	57.7	60.2	59.9	**60.3**
AR_L	70.2	72.1	71.8	**72.3**

Table 5. Ablation study of the spatial and channel attention mask.

4.6.3 Sensitivity study of global distillation

In global distillation, we rebuild the relation between different pixels to compensate for the missing global information in focal distillation and transfer it from the teacher detector to the student detector. In this subsection, we distill the student just using the global distillation with GcBlock [2] or Non-local module [29] on Faster RCNN, which is shown in Tab. 6. The results show both two relation methods extract effective global information and bring the student effective improvement, especially the GcBlock which brings 3.1 mAP improvement.

Methods	mAP	AP_S	AP_M	AP_L
baseline	38.4	21.5	42.1	50.3
Non-Local	39.8	22.7	43.1	52.3
GcBlock	**41.5**	**23.4**	**46.0**	**55.3**

Table 6. Comparison of different global relation methods on Faster RCNN ResNeXt101-Res50. Here we train the student just with global distillation.

T	0.3	0.5	0.8	1.0	1.2
mAP	40.1	**40.4**	**40.4**	40.2	40.0
mAR	56.4	**56.7**	56.6	56.5	56.4

Table 7. Ablation study of temperature hyper-parameter T on RetinaNet ResNeXt101-Res50.

4.6.4 Sensitivity study of T

In Eq. (7) and Eq. (8), we use the temperature hyper-parameter T to adjust the pixels and channels distribution of the feature map. The gap between pixels and channels becomes wider and smaller when $T < 1$ and $T > 1$, respectively. Here we conduct several experiments to investigate the influence of T. As shown in Tab. 7, when $T = 0.5$, the student gains 0.2 mAP and 0.2 mAR improvement compared with $T = 1$, which means distillation without distribution adjustment. With $T = 0.5$, the pixels and channels of high value are emphasized more and this helps the student detector focus on such crucial parts more and perform better. It is also observed the worst result is just a 0.4 mAP drop compared with the best result, indicating our method is not sensitive to the hyper-parameter T.

5. Conclusion

In this paper, we point out the student detector needs to pay attention to both the crucial parts and global relations from the teacher. Then we propose Focal and Global Distillation (FGD) to guide the student detectors. Extensive experiments on various detectors prove that our method is simple and efficient. Furthermore, our method is just based on the feature so that FGD can be applied to two-stage detectors, anchor-based one-stage detectors, and anchor-free one-stage detectors easily. The analysis shows that the student gets a really similar feature with the teacher and initializing the student with the teacher's parameters can bring another improvement. However, our understanding of how to get a better head is preliminary and left as future works.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 4, 8

[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 2

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[5] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2021. 3, 5, 6

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019. 2

[7] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021. 2, 3

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2, 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6

[10] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019. 1, 2

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 4

[12] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3588–3597, 2018. 2, 4

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[14] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. *arXiv preprint arXiv:2110.12724*, 2021. 5

[15] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 6356–6364, 2017. 2

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 5

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 5

[21] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1, 2, 5

[23] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[25] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *arXiv preprint arXiv:2006.13108*, 2020. 2, 3

[26] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2, 5

[27] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019. 1, 2

[28] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019. 2, 3, 5, 6

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2, 3, 4, 8

[30] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4

[31] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6

[32] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9657–9666, 2019. 2, 5

[33] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 1, 2

[34] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 1, 2, 4

[35] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020. 3, 6